

DISK ARRAY DEVICE

Patent Number: JP7200190
Publication date: 1995-08-04
Inventor(s): TAKAI RIICHI
Applicant(s): NEC CORP
Requested Patent: ☐ JP7200190
Application Number: JP19930355494 19931229
Priority Number(s):
IPC Classification: G06F3/06 ; G06F12/08
EC Classification:
Equivalents: JP2570614B2

Abstract

PURPOSE: To prevent the read performance from decreasing if one disk gets out of order in a disk array which stores data decentralized and stored on plural disks and their parity information.
CONSTITUTION: A cache memory 300 which holds data of respective disk drives 401-40N normally functions as a virtual disk which holds the contents of a faulty disk drive if the fault of one of the disk drives 401-40N is detected.

Data supplied from the esp@cenet database - I2

RECEIVED
JAN 10 2001
Technology Center 2100

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平7-200190

(43) 公開日 平成7年 (1995) 8月4日

(51) Int. Cl. ⁶	識別記号	庁内整理番号	F I	技術表示箇所
G 0 6 F 3/06	5 4 0			
	3 0 5 C			
12/08	3 2 0	7608-5B		

審査請求 有 請求項の数 2 書面 (全 6 頁)

(21) 出願番号	特願平5-355494
(22) 出願日	平成5年 (1993) 12月29日

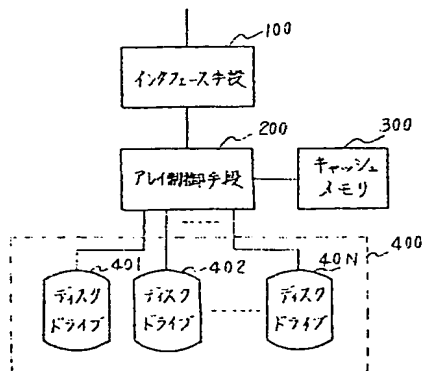
(71) 出願人	000004237 日本電気株式会社 東京都港区芝五丁目7番1号
(72) 発明者	高井 利一 東京都港区芝五丁目7番1号 日本電気株式 会社内
(74) 代理人	弁理士 京本 直樹 (外2名)

(54) 【発明の名称】 ディスクアレイ装置

(57) 【要約】

【目的】 複数のディスクに分散して格納したデータ及びそれらのパリティ情報を保存するディスクアレイにおいて、1台のディスクが故障した際の読み出し性能低下を防止する。

【構成】 正常時にはディスクドライブ401~40Nのそれぞれのデータを保持するキャッシュメモリ300は、ディスクドライブ401~40Nのいずれかにおいて故障が検出されると、その故障したディスクドライブの内容を保持する仮想ディスクとしても機能する。



【特許請求の範囲】

【請求項1】複数のディスクドライブを有するディスクドライブ群と、

これらディスクドライブ群への書き込みデータに関するエラー修復のための情報としてエラー修復情報を生成し、前記ディスクドライブ群からの読出しデータに関してエラーの発生をチェックしてエラー修復をするアレイ制御手段と、

前記複数のディスクドライブの1つが故障した場合にはこの故障したディスクドライブが格納すべきデータとして前記アレイ制御手段により修復されたデータを格納するキャッシュメモリ手段とを含むことを特徴とするディスクアレイ装置。

【請求項2】前記キャッシュメモリ手段は前記故障したディスク以外のディスクが格納するデータをも格納することを特徴とする請求項1に記載のディスクアレイ装置。

【発明の詳細な説明】

【0001】

【産業上の利用分野】本発明は、情報処理装置に於いて用いる磁気ディスクを用いた冗長性を持つディスクアレイに関し、特に故障したディスクが存在する場合のデータアクセス性能の向上に関する。

【従来の技術】従来情報処理装置において用いられる外部記憶装置はシステム側の必要によりより大容量の記憶装置を用いる方向に進んでいたが、この大容量の外部記憶装置を使用するシステム構成では1つの記憶装置の故障により大量のデータが一度に喪失してしまう。また、同一外部記憶装置の別領域に記憶されているデータへの並列アクセスが出来ないため、システム全体としてのスループットが低下してしまう。上記のような単一の大容量外部記憶装置を用いた場合の弊害をなくすために考え出された方式がカリフォルニア大学バークレー校コンピュータ科学学科により提唱された“低価格ディスクによる冗長アレイ (Redundant Arrays of Inexpensive Disk、以下RAIDという)”であり、例えば、「A Case for Redundant Array of Inexpensive Disks (RAID)」、カリフォルニア大学バークレー校コンピュータ科学学科 (EESC)、レポート番号UCB/CSD/87/391」に記載されている。このRAIDについてはRAID1よりRAID5までの5種類の実現方法が提唱されているが、このうちRAID4及びRAID5として規定されている方式ではRAIDアレイに対するデータをブロック単位で分割し、その分割したデータをアレイを構成する各ドライブに分散させて記憶し、そのパリティデータを別ドライブに記憶している。このRAID4、5に対し書き込みを行なう場合、該当データとそれに対応するパリティデータを一度読みだしてから再書き込みを行な

う必要がある。この為通常の単体ドライブに対するデータ書き込みに比べ余分なディスクアクセスが必要となる。この読み出し及び再書き込みのために必要となるオーバーヘッドをライトペナルティと呼ぶ。このライトペナルティを少なくするため通常はホストからの書き込みデータがアレイを構成している磁気ディスクへ書き込まれる前にキャッシュメモリに一度データを保持し、実際に磁気ディスクへのデータ書き込みが終了する前にアレイとしてホストへ書き込み終了を通知することにより書き込み時の性能低下を防いでいる。

【0002】また、現在実用化されているディスクアレイにおいて1台のドライブが故障した場合、他のドライブのデータより故障したドライブのデータを修復する必要があるためアレイ全体の性能が低下する。これを防ぐためには故障したドライブに記憶されているはずのデータと同等のデータを持つ仮想ドライブを用意する必要がある。

【0003】

【発明が解決しようとする課題】上述したようにディスクアレイでは自己を構成する磁気ディスク装置が故障した縮退状態に陥っている場合、故障した磁気ディスク装置に記憶されているデータを他の磁気ディスク装置より再構築する必要があるためどうしても通常時に比べ読み出し速度が低下してしまう。

【0004】本発明はディスクアレイの書き込み性能の向上のために用意されているライトキャッシュメモリを故障した磁気ディスク装置とデータの的に同等な仮想ディスクとして用いることにより縮退時のデータアクセス性能の低下を抑えることを目的とする。

【0005】

【課題を解決するための手段】上記課題を解決するため、本願発明のディスクアレイ装置では、複数のディスクドライブを有するディスクドライブ群と、これらディスクドライブ群への書き込みデータに関するエラー修復のための情報としてエラー修復情報を生成し、前記ディスクドライブ群からの読出しデータに関してエラーの発生をチェックしてエラー修復をするアレイ制御手段と、前記複数のディスクドライブの1つが故障した場合にはこの故障したディスクドライブが格納すべきデータとして前記アレイ制御手段により修復されたデータを格納するキャッシュメモリ手段とを有している。

【0006】また、前記キャッシュメモリ手段は前記故障したディスク以外のディスクが格納するデータをも格納する。

【0007】

【実施例】次に本願発明のディスクアレイ装置の一実施例について図面を参照して詳細に説明する。

【0008】図1を参照すると、本願発明の一実施例であるディスクアレイ装置は、インタフェース手段100と、アレイ制御手段200と、キャッシュメモリ300

と、ディスクドライブ群400とを含んでいる。

【0009】インタフェース手段100は、他の装置からアクセス要求を受け、また、返答を返す。

【0010】アレイ制御手段200は、ディスクアレイ装置全体の制御を行なう。すなわち、ディスクドライブ群400にアクセスすると共に、キャッシュメモリ300の管理を行なう。

【0011】キャッシュメモリ300は、ディスクドライブ群400の内容を一時的に格納するディスクキャッシュの機能を有し、また、後述するように故障したディスクに相当するデータを格納する。

$$D40N = D401 \text{ xor } D402 \text{ xor } D403 \text{ xor } \dots \text{ xor } D40(N-1) \quad \dots \text{式(1)}$$

で表されるものとする。但し、“xor”は排他的論理和を示し、 $D40x$ はそれぞれディスク40xの対応するデータを示す。ディスク40Nがこの様に定義された場合、任意のディスクに保持されているデータは他の

(N-1) 台のディスクに保持されているデータの排他的論理和と等しくなるため、1台のディスクが故障した場合でも他の(N-1) 台のディスクのデータを読み出しそのデータの排他的論理和を求めればデータ復旧を行うことが可能となる。通常、ディスクとのデータアクセスはセクタと呼ばれるあらかじめ決められたデータサイズを単位として行われるが、RAIDアレイに対しセク

$$N_{\text{new}} = (A_{\text{old}} \text{ xor } A_{\text{new}}) \text{ xor } N_{\text{old}}$$

であらわすことができることを利用し、パリティディスク40Nとディスク40Aに対する読み出しと書き込みだけで新たなデータの書き込みを終了できる。但し、この処理は単独のディスクに対する書き込みならば1回行なえば済んでいたディスクアクセスが、ディスクアレイでは書き込み処理中に、(1)パリティディスク40NからのNo1dの読み出し、(2)データディスク40AからのAo1dの読み出し、(3)パリティディスク40NへのNnewの書き出し、(4)データディスク40AへのAnewの書き出し、の4回のアクセスを行なう必要が生じるため、1回の書き込み終了時間が遅くなる。以下、これをライトペナルティとよぶ。

【0014】本願発明のディスクアレイ装置の一実施例

$$D40X = D401 \text{ xor } \dots D40(X-1) \text{ xor } D40(X+1) \dots \text{ xor } D40N \quad \dots \text{式(3)}$$

と言う関係が成り立つため、他の(N-1) 台のディスクデータの排他的論理和を計算することで、故障ディスク40Xのデータを復旧することが可能となる。なお、このようにディスクの1台が故障し、他のディスクのデータからデータを計算により算出している状態を縮退状態と呼ぶ。一般に、この縮退状態では通常は40Xのディスクへのアクセスですむデータ読み出しを行う代わりに、40Xを除く(N-1) 台のディスクへのアクセスが必要となるため、正常状態よりもアクセス性能が落ちる。この性能低下を防止するため、本願発明のディスクアレイ装置の一実施例では、縮退状態においては、キャッシュメモリ300が故障ディスク40Xに相当する内

【0012】ディスクドライブ群400は、さらにN台のディスクドライブ401~40Nを含んでいる。RAID構成をとるディスクアレイの場合、上記N台のディスクのうちの1台をパリティデータ保持に使用する。本実施例においてはディスク40Nをパリティデータ保持に使用するものとする。

【0013】本願発明のディスクアレイ装置の一実施例では、通常のRAID構成を持つディスクアレイ装置と同様に、パリティデータ保持に使用されるディスク40Nに書き込まれるデータD40Nは、

タの大きさ以下の書き込みを行う場合には、複数のディスクへの同時アクセスを可能とするために、書き込み時に全てのディスクのデータを書き換えるのではなく、書き込みデータの保持を行なうディスク40Aとパリティディスク40Nにのみアクセスを行う必要がある。すなわち、ディスク40Aに既に書かれているデータをAo1d、これから書き込まれるデータをAnewとし、パリティディスク40Nに書かれていたデータをNo1dとした場合、パリティディスクに新たに書き込む必要のあるデータNnewは他のディスクに書き込まれているデータには関係なく、

$$\dots \text{式(2)}$$

では、通常のRAID構成を持つディスクアレイ装置と同様に、このライトペナルティを短縮するために、一時的にキャッシュメモリ300に書き込みデータを保持し、実際にディスク401、402...40Nにデータが書き込まれていなくともキャッシュメモリ300にデータが書き込まれた時点で、インタフェース手段100に対し終了報告を行う方式を採用する。

【0015】このように構成することにより、任意のディスク40Xが故障した場合(以下、この故障したディスク40Xを故障ディスクという)、通常の単独ディスクではそのディスク40Xに格納されていたデータを復旧することは不可能であるが、RAID構成を持つアレイでは式(1)より、

容を格納することにより、故障ディスク40Xへのアクセス時に復旧をしないで済むことになる。

【0016】図2(A)を参照すると、本発明の一実施例のディスクアレイ装置が正常なディスクアレイとして動作している場合のキャッシュメモリ300のメモリマップにおいて、キャッシュメモリ300は各ディスク501、502...50Nに書き込まれるべきデータを一時的に保持している。インタフェース手段100より転送された書き込みデータはディスク501、502...50Nへの書き込み前に各々のディスクに対応させたディスクキャッシュ301、302...30Nに一度書き込まれ、この段階でインタフェース手段100を介して、上

位装置に対し終了報告を行うことにより物理動作が必要なディスク401、402…40Nへデータを書き込む場合に比べ、システム全体の書き込み性能の向上を図れる。

【0017】図2(B)を参照すると、縮退状態のキャッシュメモリ300のメモリマップにおいて、キャッシュメモリ300は書き込みデータ用のディスクキャッシュ301、302…30Nとしてではなく、故障したディスク40Xと等価な仮想ディスク700として位置付けられ、故障ディスク40Xに対するデータ読み出し時に再構成されたデータを仮想ディスク700に記憶させる。つまり、正常時はアレイに書き込まれたデータを保存し次に書き込まれたデータを読み出す場合にメモリ上のデータを使用することにより1回のディスクアクセスを無くしているが、縮退時には故障しているディスク40Xのデータを式(3)により再構成した結果を保存するために使用することにより故障ディスクデータの読み出し時に必要な(N-1)台のディスクへのアクセスを減らすことを可能とする。この様に一度読み出し処理が行われた事により再構成された故障ディスク40Xと等価なデータを仮想ディスク700に蓄積し、同一領域を再度読み出す場合にはこの蓄積されているデータを読み出しデータとして使用することにより縮退状態における読み出し性能の低下を防ぐことが可能となる。この場合、キャッシュメモリ300の容量はディスク401、402…40Nと同等以上であることが望ましいが、ディスクより容量が少ない場合でも、通常のキャッシュメモリと同様の機能が働き、使用頻度の高いデータを優先して保持することになるため、仮想ディスク700に記憶されているデータのヒット率は一定のレベルに保つことが可能である。

【0018】次に本願発明の一実施例のディスクアレイ装置の動作について説明する。

【0019】図3を参照すると、本願発明の一実施例のディスクアレイ装置の縮退時の読出し動作を表す処理の流れ図である。上位装置からのアクセス要求はインタフェース手段100を介してアレイ制御手段200に渡され、このアレイ制御手段200で解読される。このアクセス要求が読み出し要求であれば、アレイ制御手段200は図3の流れ図に従って、故障ディスク40Xに対するアクセスが否かが判断される(ステップ201)。故障ディスク40X以外のディスクドライブに対する読み出し要求であれば、アレイ制御手段200は該当するディスクドライブから読み出しを行なう(ステップ202)。

【0020】故障ディスク40Xに対する読み出し要求であれば(ステップ201)、当該データがキャッシュメモリ300に格納されているか否かが調べられる(ステップ203)。当該データがキャッシュメモリ300に格納されていれば、すなわちキャッシュヒットであれ

ば、当該データがキャッシュメモリ300から読み出される(ステップ204)。

【0021】故障ディスク40Xに対する読み出し要求であって、しかも当該データがキャッシュメモリ300に格納されていなければ(ステップ203)、アレイ制御手段200はディスクドライブ群400から当該アクセス要求に対応するパリティ、すなわち式(1)が成立するD40N、をパリティディスク40Nから読み出す(ステップ205)。

10 【0022】続いて、アレイ制御手段200は当該アクセス要求に係るデータに対応するデータ群、すなわち式(1)が成立するD401~D40(N-1)、をディスクドライブ群400から読み出す(ステップ206)。但し、故障ディスク40Xからはデータを読み出さない。

20 【0023】そして、上記パリティD40Nとデータ群D401~D40(N-1)とから、式(3)によって、故障ディスク40Xに対応するデータD40Xを復旧する(ステップ207)。このようにして復旧されたデータD40Xは、キャッシュメモリ300に書き込まれる(ステップ208)。

【0024】図4を参照すると、本願発明の一実施例のディスクアレイ装置の縮退時の書き込み動作を表す処理の流れ図である。上位装置からのアクセス要求が書き込み要求であれば、アレイ制御手段200は図4の流れ図に従って、ディスクドライブ群400から当該アクセス要求に対応するパリティ、すなわち式(1)が成立するD40N、をパリティディスク40Nから読み出す(ステップ211)。

30 【0025】続いて、アレイ制御手段200は当該アクセス要求に係るデータに対応するデータ群、すなわち式(1)が成立するD401~D40(N-1)、をディスクドライブ群400から読み出す(ステップ212)。但し、故障ディスク40Xからはデータを読み出さない。

40 【0026】そして、アレイ制御手段200は書き込みデータを新たに含めたデータ群D401~D40(N-1)から式(1)によってパリティD40Nを生成して、このパリティD40Nをパリティディスク40Nに書き込む(ステップ213)。

【0027】その後、書き込みデータが故障ディスク40Xに対するものであればこの書き込みデータはキャッシュメモリ300に書き込まれ(ステップ216)、故障ディスク40X以外のディスクドライブに対するものであれば当該ディスクドライブに書き込まれる(ステップ215)。

50 【0028】図5及び図6を参照すると、図5はキャッシュメモリ300をディスクキャッシュとして用いた場合のデータ転送ルートを示す図であり、図6は縮退時にキャッシュメモリ300を仮想ディスク700として使

用した場合のデータ転送ルートを示す図である。

【0029】なお、縮退時にキャッシュメモリ300を仮想ディスク700として用いた場合、前に示したライトペナルティを避ける方法が無くなるためにより書き込み性能は低下する。この性能低下は、図2(C)にメモリマップを示したように縮退時において常にキャッシュメモリ300の全容量を仮想ディスク700として確保するのではなく、故障ディスクと等価なディスクとして再構成されたデータを保存していない部分をディスクキャッシュとして用いることにより防ぐことが可能となる。つまり、キャッシュメモリ300上にライトペナルティを防ぐキャッシュとして利用する領域301~30Nと仮想ディスク700として利用する領域を双方用意することにより、仮想ディスク領域700で縮退時の読み出し性能の低下を、キャッシュ領域301~30Nでライトペナルティの防止を行なうことが可能となる。

【0030】

【発明の効果】以上説明したように、本発明はディスクアレイにおけるライトペナルティの補完用であるキャッシュメモリを、縮退した磁気ディスク装置の仮想ディスクとして用いることにより、縮退したディスクアレイからのデータ読み出し時の性能低下を抑えることを可能と

する効果を有する。

【図面の簡単な説明】

【図1】本発明のディスクアレイ装置の一実施例を表すブロック図である。

【図2】本発明の一実施例のキャッシュメモリが格納するデータのメモリマップを表す図である。

【図3】本発明のディスクアレイ装置の一実施例の読出し動作の処理の流れを表す図である。

【図4】本発明のディスクアレイ装置の一実施例の書き込み動作の処理の流れを表す図である。

【図5】本発明のディスクアレイ装置の一実施例における正常時のデータ転送ルートを表す図である。

【図6】本発明のディスクアレイ装置の一実施例における故障発生時のデータ転送ルートを表す図である。

【符号の説明】

100 インタフェース手段

200 アレイ制御手段

300 キャッシュメモリ

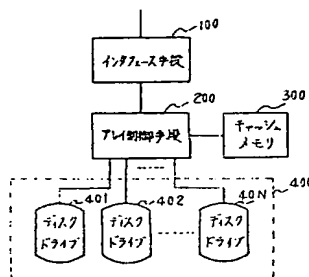
301~30N ディスクキャッシュ領域

400 ディスクドライブ群

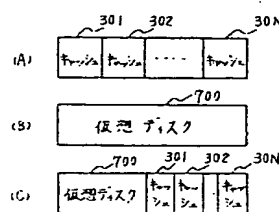
401~40N ディスクドライブ

700 仮想メモリ領域

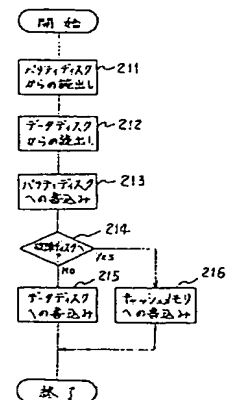
【図1】



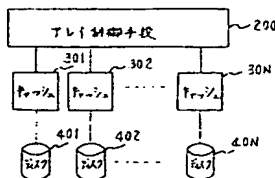
【図2】



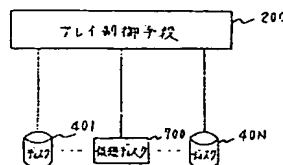
【図4】



【図5】



【図6】



【図3】

